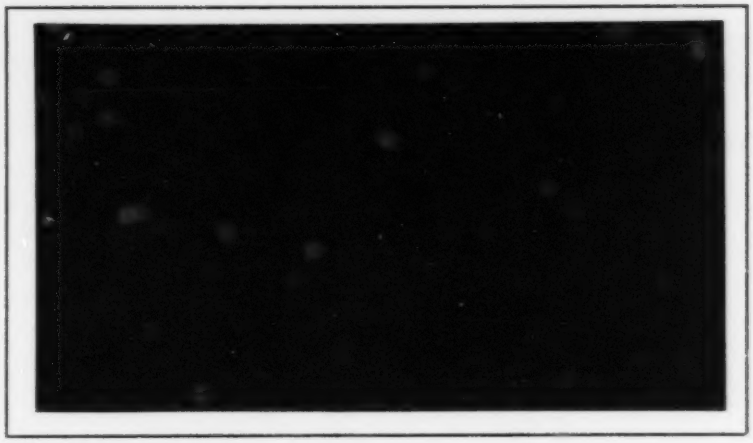

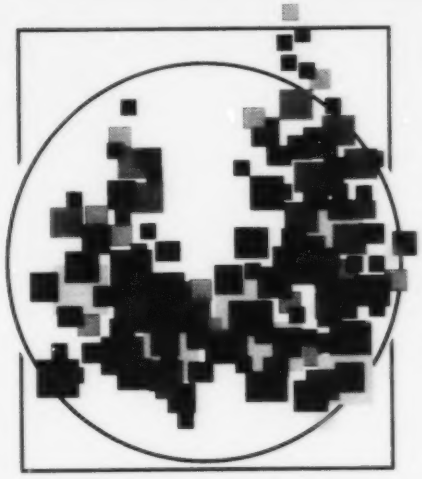

CENTRE FOR
SYSTEMS
SCIENCE

Simon Fraser University, Burnaby, B.C., Canada V5A 1S6 / (604) 291-4704

92-05736/

Connectionism, Rule Following, and Symbolic Manipulation

Robert F. Hadley

CSS-IS TR 90-03

March 1990

Connectionism, Rule Following, and Symbolic Manipulation

by

Robert F. Hadley

School of Computing Science
Simon Fraser University
Burnaby, Canada V5A 1S6
hadley@cs.sfu.ca

Abstract

At present, the prevailing Connectionist methodology for *representing rules* is to *implicitly* embody rules in "neurally-wired" networks. That is, the methodology adopts the stance that rules must either be hard-wired or "trained into" neural structures, rather than represented via explicit symbolic structures. Even recent attempts to implement *production systems* within connectionist networks have assumed that condition-action rules (or rule schema) are to be embodied in the *structure* of individual networks. Such networks must be grown or trained over a significant span of time. However, arguments are presented herein that humans *sometimes* follow rules which are *very rapidly* assigned *explicit* internal representations, and that humans possess *general* mechanisms capable of interpreting and following such rules. In particular, arguments are presented that the *speed* with which humans are able to follow rules of *novel structure* demonstrates the existence of general-purpose rule following mechanisms. It is further argued that the existence of general-purpose rule following mechanisms strongly indicates that explicit rule following is not an *isolated* phenomenon, but may well be a pervasive aspect of cognition. The arguments presented here are pragmatic in nature, and are contrasted with the *kind* of arguments developed by Fodor and Pylyshyn in their recent, influential paper.

1. Introduction

In a recent and influential paper, Fodor and Pylyshyn present principled arguments to the effect that widespread methods of representation in connectionist networks are incapable of manifesting certain properties which occur as essential aspects of human cognition. These include compositionality, generalizability, and systematicity, all of which are prevalent in human thought and language. The kernel of Fodor's and Pylyshyn's position is that these crucial properties presuppose the existence of structure-sensitive operations which, of necessity, occur at a higher level of abstraction than that which is typically associated with connectionist processing. Moreover, the required structure-sensitive operations presuppose *structured representations* which do not exist at the level of the local and distributed representation schemes prevalent in the connectionist literature (or so it is argued). Now, although some connectionists remain skeptical about Fodor and Pylyshyn's ultimate conclusions, many concede that compositionality, generalizability, and systematicity are indeed central aspects of human

cognition which connectionism, as a general enterprise, must address.

Recently, Elman (1989) and St. John & McClelland (1989) have devised connectionist networks (hereafter, c-nets) which exhibit these crucial properties, while remaining faithful to conventional (distributed) methods of connectionist representation. Several intriguing issues are raised by these results, which I shall touch upon only obliquely. My primary concern shall be to describe certain human cognitive abilities which challenge the connectionist thesis on grounds different from those put forth by Fodor and Pylyshyn. In particular, I describe cognitive abilities which involve (virtually) instantaneous rule learning and *application* of these rules to data which are retained in short term memory. Such abilities cast doubt upon the widespread connectionist practice of modelling the acquisition of *all general* rules by the training (or hard-wiring) of c-nets. While I do not question whether c-nets could *eventually* be trained to display the relevant cognitive behavior, I argue that the *speed* with which humans are able to acquire and follow rules of novel structure demonstrates both that humans *sometimes* represent rules in an explicit fashion (in a sense of 'explicit' defined below), and that they possess *general-purpose* mechanisms for applying such rules. Moreover, one of the examples presented here involves such conscious and explicit rule following and symbol manipulation that if *essentially* connectionist mechanisms are involved, we seem forced to conclude that sometimes, at least, c-nets merely provide the architectural foundation for conventional, structure-sensitive, symbol manipulation.

2. (Nearly) Instantaneous Rule Acquisition

In this section we examine a methodological principle that is frequently assumed (and sometimes fervently believed) by connectionists, namely, that cognitive processes which are ostensibly *rule-governed* ought to be modelled on the assumption that *individual rules* are *embedded* in the structure or weight distribution of particular c-nets, and should not be modelled as symbolic code which is applied to data sets on different occasions. Against this approach, I shall argue that *if all* (human) higher-level cognitive processes can be modelled by a system of c-nets, then we must suppose that at least *some* of these c-nets function as *general purpose* rule-interpreters which can *apply* rules to arbitrary sets of input. In particular, I argue that *some* rule-like behavior cannot be the product of "neurally-wired" rules whose structure is embedded in particular networks, for the simple reason that humans can often apply rules (with considerable accuracy) as soon as they are told the rules.[1] The following example will help to illustrate this. Consider the phrase:

Example 1: "love ever keeps trying".

While holding this phrase in mind, try applying the rule described in the footnote indicated at this point.[2] If you succeeded in remembering the given phrase while you applied the indicated rule (which is italicized), then you discovered that this rule, applied to this datum, yields the word 'over'. (Otherwise, you were no doubt distracted by having to read the surrounding instructions, jumping to footnotes, etc.) Most people have no difficulty finding the correct answer when the experiment is verbally described, face to face.

Now, the immediate thing to note about this experiment is that we can promptly and correctly apply the rule to a mentally retained datum, even though we have never encountered the rule or the datum before. The fact that we can *comprehend* novel rules and phrases argues for the compositionality and systematicity of thought, but *that* is not my point here. (In any case, recent work by St. John & McClelland, 1989, demonstrates that, at least in some experimental conditions, c-nets can develop and represent "semantic interpretations" of novel sentences.) Rather, the point is that we have never been *trained to respond* either to this particular rule, or to this datum, or to their joint occurrence. Nevertheless, we are not only able to comprehend the rule, but to *act in accordance* with it. Given that we are able to follow the rule immediately, it would be wild to suppose that, in the short time available, the rule is somehow transformed into an appropriately trained network, which (implicitly) implements this *particular* rule. (Bear in

---

[1]The point here is not simply that some behavior is "cognitively penetrable" (in Pylyshyn's sense, cf. Pylyshyn, 1984, p. 133). For, as Pylyshyn notes, the fact that behavior is cognitively penetrable is explained by the fact that it is rule-governed. But even the prevailing connectionist methods of representing rules are compatible with supposing that much of our behavior is rule-governed. The question I address is whether all rule-governed behavior can be explained by the methods for instantiating rules now advocated by connectionists.

[2]The phrase you have in mind contains four words. *Proceeding from left to right, mentally extract the second letter from each word, and concatenate these letters in sequence. If the resulting string forms an English word, make a note of it.*

mind that our example rule is *general*, in the sense that it may be applied to many distinct inputs. Widespread experience has established that rules of this degree of generality can be trained into c-nets **only** via gradual tuning of weights, involving many small incremental changes.) It appears, therefore, that we can safely rule out the possibility that our brains contain a c-net which implicitly embodies this specific rule (via, hard-wiring or a distributed set of weights, as the conventional connectionist paradigm would suggest).

The question remains, then, how *could* this novel rule be rapidly executed by a system of one or more c-nets? Well, if the rule is being executed by c-net(s), and no c-net is specifically trained for the rule, we can only suppose that once the rule has been processed as sensory input (and perhaps been assigned an internal representation (local or distributed)), the rule functions as *data* which causes some fairly *general purpose* c-net(s) to do just what the rule tells us to do. The c-net(s) which "react" to the rule being input must, in some sense, be *fairly general*, because we have already seen the implausibility of postulating c-nets which are specific to each novel rule that we can execute. However, in saying the c-nets are "fairly general", I do not preclude the possibility that very different kinds of c-nets might be required to process radically different kinds of rules. At this point, the *degree* of generality of c-nets which function as *rule-interpreters* and *executors* must be left partially indeterminate. It seems likely, however, that we could invent a whole family of rules of the same type as the one we have considered, and that each of these rules would be processed by the same set of c-nets (up to a point, at least. Ultimately, individual words, say 'two' vs. 'three', would presumably involve different subnets.)

Now, at this point the following objection may arise:

> Well, yes, given that we *sometimes interpret* and *apply* novel rules, and given that these rules cannot plausibly be supposed to be (innately, or by training) neurally-wired in our brains, then we *must* suppose that the brain contains sets of c-nets which function as (moderately) general rule interpreters, provided the brain is a collection of c-nets. But, note that the example you describe is a-typical, and does not establish that rule interpretation and application are important or common processes in human cognition.

Before replying to the above objection, it will be helpful to introduce some terminology. Hereafter, we shall say that an *explicit* rule has been followed, if and only if a rule has been followed, but the rule is *not embedded* in the structure or weight distribution of some c-net. (I am here using 'explicit' as a technical term. No claim is made that this technical usage corresponds closely to common usage. Admittedly, the natural language distinction between 'implicit' and 'explicit' is slippery and difficult to unravel. However, our present concern is to distinguish rules which are embedded in the structure or weight-distribution of c-nets from those which are not.) Also, it is important to realize that example (1) establishes *not only* that we sometimes follow rules which are explicit in the sense just defined, but that some of these explicit rules are internally represented when we follow them. To see this, consider that in example (1) an explicit (non-embedded) internal representation of the rule must be posited to explain the fact that subjects are able to follow the rule for several minutes after it is spoken,

even when other, extraneous utterances intervene.[3]

Now, to reply, first note that *even if* the interpretation and application of non-embedded (explicit) rules are uncommon cognitive events, they do occur, and their explanation seems to require a departure from the methodological principle described at the beginning of section 2. In addition, the example establishes that it is not only a theoretical possibility that neural structures could support the *general* application of *explicit* rules, but this is sometimes a reality. It is important that connectionists bear in mind that at least *some* of our neural structures are organized in ways that permit the use of rules that have not been trained (or grown) into a net. Moreover, we need to consider how it happens, if explicit rule use is such an a-typical cognitive phenomena, that we *possess* an assembly of c-nets capable of performing such feats. There appear to be only two possibilities: either the c-net(s) *primarily* responsible for the rapid interpretation and application of rules are *innately* present in the brain (with the appropriate weights pre-set), or they are not. Let us consider these possibilities in turn.

Suppose the relevant set of c-nets to be innately hard-wired. Now, it is no doubt a very difficult problem to distinguish the relevant set of *innately* hard-wired c-nets (if they do exist) from *other* c-nets involved in language understanding (which must be *trained* during language acquisition), but fortunately we may sidestep that problem. For the point is that if these innate c-nets are *primarily* responsible for our general ability to rapidly apply novel rules, then the need for such *general capacity* c-nets must arise rather often. Otherwise, (a) it is unlikely that the relevant c-nets would have evolved in the first place, and (b) even supposing that these c-nets had *not* evolved specifically to handle rule application, it is not plausible that they should integrate so rapidly and accurately with our general language comprehension mechanisms that we would be able to apply novel rules with the facility that we exhibit. I conclude that if the relevant set of c-nets are innately hard-wired, this strongly suggests that *explicit* rule application is not a rare event, but is an important (and probably common) aspect of our cognitive life.

On the other hand, suppose the relevant c-nets are *not innately* given (with pre-sets weights). In this case we must suppose the relevant c-nets are either *specifically trained* to perform the *general* task of rule application, or the c-nets possess these abilities as a side-effect of other abilities. In the former case we cannot suppose the c-nets in question would receive the required specialized training, unless the general task of *applying* rules to representations was frequently encountered. So, in this case the ability to apply explicit rules can hardly be regarded as an

---

[3]The objection may arise that we need not suppose that a representation of the rule is stored beyond the first few seconds, for it is logically possible that a c-net will rapidly be trained to *implicitly* embody the rule once the rule has been applied to the *first* input set. However, this objection presumes the existence of biological mechanisms which are able, very rapidly, to train up a network to perform a *general* task which has only once been comprehended and performed. Apart from the sheer physical implausibility that the weights of individual axons could be sufficiently modified in the available time (changing the weight of a single axon requires growing cells to augment the axon's diameter), the objection ignores the fact that c-nets can only be trained to acquire *general* rules by a gradual iterative process.

isolated phenomenon, as the objector implies. Let us consider, therefore, the latter case, in which the ability to apply novel rules arises as a *side-effect* of other abilities.

We should note at the outset that the ability to apply novel rules may often, and perhaps always, involve a series of *sub-skills* which have been acquired through slow learning. For example, the rule I presented earlier may involve the sub-skills of retrieving the spelling of a word, of selecting the n-th element in a list (in this case a list of letters), and of concatenating letters to form a word. For argument sake, I concede that each of these skills may have been acquired by slow, iterative training of c-nets. However, the mere presence of these separate skills does not explain how we should be able, as a *side-effect* of these abilities, to create a *coherent sequence* of operations, in which each skill is invoked at the proper time, and applied to the proper object. By analogy, the mere presence of a set of primitive operations in a programming language does not cause a program to be written, assembled, or executed. In short, the existence of primitive, slowly learned skills in a neural system may be a necessary condition for the application of novel rules, but it is *not* a *sufficient* condition. If we are to preserve the "side-effect hypothesis", then the relevant side-effects must arise from c-nets other than (or in addition to) those responsible for executing the sub-skills involved.

Now, while I know of no way to disprove this possibility, it does seem odd that such a complex ability as explicit rule following would arise as a mere side-effect of other neural processes. In any case, there is a deeper point to be made here. For, even if the side-effect hypothesis is correct, connectionism has not provided us with any reason for supposing that side-effects of this kind are limited in their scope. On the contrary, if side-effects of collections of c-nets are capable of supporting *rapid* application of completely novel rules, why should we not suppose that much, or even most higher-level cognition also involves the explicit application of rules which are acquired through direct observation or through explicit teaching? Why should we not suppose that *many* rules are stored in non-embedded form, and are interpreted as the need arises? Such a hypothesis would not exclude the *further* conjecture that when a rule enters *long-term* memory, some c-net will be trained to *implicitly* represent the rule by means of acquired weights. However, since neural mechanisms are clearly capable of applying explicit rules, we should await clear empirical evidence before judging the pervasiveness of this form of rule following. In sections 4 and 5, I present two examples which suggest that explicit rule following is more pervasive than a connectionist might suppose, but before passing to these examples, let's consider what more we can learn from the present example.

3. Manipulation of Symbolic Representations

Recall that in example (1) the subject is asked to keep a phrase in mind, in this case "love ever keeps trying". This phrase must be retained (presumably, in short term memory, or some other buffer region) while the subject listens to a rule. After hearing the rule the subject *somehow* retrieves the individual words of the given phrase, in sequence, in order to select the second letter from each word. At least, this is how it appears to us introspectively. However, from a purely logical standpoint, we need not suppose that individual words (or representations of

words) are being reviewed in sequence. We may choose to ignore introspective evidence (though such evidence seems to require *some* explanation), and suppose that the input phrase is assigned an internal representation which is *not spatially* composite.[4] In what immediately follows, we will accept this supposition, since it appears to represent the "worst case" for what I wish to demonstrate, viz., that example (1) involves explicit (mental) symbol manipulation.

Now, if our mental representation of the example phrase is not spatially composite, then we have two possibilities. Either the phrase is internally represented by a single node (i.e., it is assigned a *local* representation) or it is assigned a *distributed* representation whose spatial parts are not themselves meaningful representations. For simplicity sake, and for reasons given by Fodor and Pylyshyn (1988) we shall not consider the localist approach.[5] So, we assume that the input phrase is assigned a distributed representation. (An account of how phrases and sentences may be assigned distributed *meaning* representations is given in St. John & McClelland, 1989).

Now, although spatial sub-regions of this distributed representation are assumed *not* to be representations of any kind, it is still conceivable that some c-net exists which, given this distributed representation, and primed with the rule in question, could simply output a representation for the word 'over', i.e., the answer word. But while this is *conceivable*, it seems rather doubtful. For, since subjects receive no training at the *particular* task in question, it is not reasonable to suppose that *any* c-net contains information (tacit, or otherwise) about the specific letters occurring at specific positions in the particular phrase or sentence being represented. Moreover, the *general* task of retrieving letters from *entire phrases* at specified positions is not one that people are commonly trained for. By contrast, the task of retrieving the spelling of individual *words* is one that we are trained for, as is the task of finding the n-th element in a series of objects (e.g., a series of letters). It is entirely plausible, therefore, that we should have c-nets capable of performing these *sub-tasks*. Now, given the complexity of the task of going from an arbitrary phrase representation to the spelling of the answer word, it would be strange indeed if the c-nets comprising the *general rule interpreter* (which we have already seen to be necessary) did not arrange for the relevant subtasks to be performed by c-nets which have already been specifically trained for those sub-tasks. Moreover, if we adopt this hypothesis -- that a series of sub-tasks are performed by c-nets specific to those tasks -- we need not abandon the idea that the input phrase receives a spatially non-composite distributed representation. For, it is plausible that we possess c-nets capable of taking this spatially non-composite representation as input, and yielding a *spatially* (or temporally) sequenced *series* of representations of the individual words in the sentence. In fact, c-nets which performed this transformation would

---

[4]We will say that a representation is *spatially composite* if some of its spatial parts are themselves meaningful representations (just as the words of this sentence are meaningful spatial parts of the entire sentence). For more on this, see (van Gelder, 1989).

[5]As Fodor and Pylyshyn stress, we cannot suppose that each phrase is represented by a unique neuron, because the number of phrases we can comprehend exceeds the number of neurons available.

simply embody a *partial inverse* of the process which produced the internal representation from the original input phrase.

The kind of transformation just described illustrates *one* way of achieving *functional* compositionality (as opposed to spatially concatenative compositionality) in connectionist architectures. Both Smolensky (1987) and van Gelder (1989) have explored the feasibility of incorporating functional compositionality in connectionist networks. Van Gelder, in particular, argues that the potential for including this form of compositionality in c-nets removes a barrier to the thesis that c-nets can exhibit the kind of systematicity which Fodor and Pylyshyn persuasively argue to be necessary. He argues further that, if connectionists are to avoid the kind of high-level control algorithms associated with *classical* AI, they will need to develop connectionist mechanisms for exploiting functional compositionality. While I do not dispute this conclusion, I contend that the present example shows that the classical paradigm of symbol manipulation is the most appropriate for some cognitive processes. For, in the absence of any plausible, *direct* c-net transformation from our example phrase to the correct answer, I submit that we *should* conclude the following: at *some* stage in the solution of our exercise the spellings of individual words are retrieved, and letters in the second position of each spelling are identified (and *in some sense* selected).

If the above conclusion is accepted, we seem committed to a process which is at least strongly analogous to classical symbol manipulation. For, consider the possibilities. Although the spellings of each of the four individual words need not be simultaneously present (in some buffer, say) still each of the four spellings must either be "consulted" in sequence or in parallel. In either case, the spellings must be examined to obtain the second letter of the given word. Now, the (representations of) letters of a given spelling must either be examined sequentially until the second element is located, or the letters are present all at once and the second letter is identified (perhaps by parallel processing). In all of the above cases we either have a (spatially) concatenated series (of letters within words) or a temporally concatenated series.[6] In either case, we have a concatenated series of representations, which taken collectively represent higher-level objects (words), and which are being processed to obtain the n-th element of the series. Moreover, once the n-th (2nd, in our case) element of each series is identified, it must somehow be marked, copied, or otherwise remembered, and its sequential position (relative to the original string of words) must be implicitly or explicitly remembered. Ultimately, these separate elements must be combined (or at least treated) as a concatenated series to obtain the representation of the external symbol 'over'.

Now, the foregoing description involves several operations which are typical of classical

---

[6]It might be argued that a third possibility exists -- that the spelling of each word is neither spatially nor temporally concatenated, but is an arbitrary (local or distributed) representation. It is conceivable that a c-net exists which takes such representations, and a parameter such as "position 2", and returns representations of individual letters. However, as we previously argued, since we rarely have training for tasks as specific as "give me the second letter of the word 'ever'", it is not plausible that such a specialized c-net would exist.

symbol manipulation (e.g., searching a list, marking or copying, selection, concatenation), but it could be argued that in one respect this description departs from classical processing. That is, I have allowed that the symbolic elements of a representation may be *temporally* rather than *spatially* concatenated. This might happen, for example, if the c-net which *functionally* decomposes a representation into its parts does so by producing those parts in a temporal sequence. But, even this kind of processing does not violate the spirit of classical symbol manipulation. Indeed, temporal concatenation seems only a minor modification (if at all) of the classical symbol manipulation paradigm. (Recall, after all, that a computer sends its symbolic output to the printer character by character.) To be sure, I have not shown that the *sequentially ordered* letters which comprise the final answer are *literally* spatially concatenated to produce this answer, but we should not expect the micro-details of how the foregoing operations are performed to resemble the micro-details of a digital simulation of these operations. For, as Pylyshyn would put it, the operations we have been considering are classical symbol manipulations at the *cognitive level of description*. There is a literal isomorphism between the series of sub-tasks performed by the c-nets involved and the moderately high-level sub-tasks involved in a computer simulation of this example. Moreover, the kinds of arguments produced earlier (to the effect that cognitive mechanisms which support explicit rule manipulation are not likely to be isolated aberrations) also apply here. We are not yet in a position to say how pervasive classical symbol manipulation is within higher-level cognition. However, the following sections may shed light upon this issue.

4. Condition-Action Rules

Consider the following simple rule:

Example 2. If I pause and say a word which ends with 'ly', then you repeat that word.

Humans are able to learn simple novel rules of this kind as soon as they are spoken. Such rules are not remembered for long (without special motivation), but they are easily retained for 15 minutes, and can be applied immediately.[7] Now, as in our previous example, we are confronted with a rule which can be immediately *retained, interpreted,* and *applied*. For reasons previously stated, we must suppose this novel rule is interpreted and applied by (at least) moderately general-purpose c-nets. Moreover, it is reasonable to believe that these c-nets receive a *representation* of the rule as input. (As in example (1), an internal representation must be posited to explain the fact that subjects can, for many minutes, reapply the rule to new data without being reminded of the rule.)

Now, unlike our previous example, the present rule has the same *general form* as other rules which we *do* learn by training. For example, "If the light turns green, then you can go" also has this general form. It is *possible* therefore, that the particular c-nets responsible for applying rule

---

[7] I've tested rules of this type on different occasions in graduate courses. Although no controlled experiment was conducted, the results were so unambiguous as to leave little doubt. People are able to follow the rule for some time after hearing it, even when ordinary conversation intervenes between "testing moments".

(2) have been trained just to recognize if-then, condition-action rules. Note, however, that condition-action rules (which usually lack truth values) structurally resemble *propositional* if-then rules (which have truth values). In light of this structural similarity, it would not be surprising if the same c-nets were responsible for interpreting and applying both *condition-action* rules (such as (2)) and for executing *modus ponens* in the presence of if-then propositions. In any case, it appears that the existence of *some* general c-nets, capable of recognizing and applying a variety of *novel* condition-action rules, should be conceded.

These c-nets may or may not be specific only to condition-action rules, but once their existence is conceded the possibility certainly arises that these c-nets are involved in widespread condition-action (or production system) reasoning. Recognition of this fact may well have motivated two recent connectionist attempts to model *production systems* (cf. Touretzky & Hinton, 1988; Ajjanagadde & Shastri, 1989). Although both these systems lack the full generality of AI-type production systems, each represents an attempt to incorporate *structure-sensitive* rule firing mechanisms in a connectionist framework. It is also worth noting that each of these systems *represent* if-then rules as *structured*, spatially composite objects, where the antecedents and consequents of individual rules are spatially separate clusters. Although these implementations do not employ classical symbol manipulation techniques at the micro-level, there is a clear and appropriate level of description at which they are performing quantifier instantiation and *modus ponens*. Unfortunately, it is doubtful whether either implementation can account for the *immediate* application of *novel* if-then rules, since in both systems, distinct conditionals are represented by distinct *neurally-wired networks*. It is not possible that such networks could be "instantly grown" to represent a rule which has just been understood. Barnden (1988) describes a method by which *constants* (which replace variables in rule schemata) could be rapidly represented in connectionist matrices, but his approach also requires the hard-wiring of each *rule schema*, and no indication is given of how *this* could be accomplished "on the fly".

5. Rules of Arbitrary Structure.

As we have noted, the previous rule (2) shares a general form with other rules we typically encounter. For this reason it is at least plausible that rules having this form are *applied* by c-net(s) which respond only to rules of this form. However, the *following* rule, like the rule in example (1), does *not* share a syntactic structure with rules we normally encounter. The background context for the following rule is this: The subject is told that she/he will be presented with a series of at most five integers, where the value of each integer is less than five, and the subject is to apply the rule (below) once the series is presented.

Example 3. Regard the second integer in the series as an exponent. Take the last integer in the series and raise it to the power indicated by the exponent you identified.

Now, college students have no difficulty understanding and applying this rule to series of integers which are presented to them. (At least the students in my sample had no difficulty. In any case, we only need a few successes to make the point.) Also, because the rule is novel,

moderately complex, and unusual in structure we must suppose, as we did in example (1), that *initial* applications of the rule involve (at least) moderately general c-nets, capable of interpreting and applying a considerable range of possible rules. Moreover, as in the preceding examples, we must assume that (temporarily at least) the rule is stored as a representation, for the rule can be repeatedly applied without being re-presented to the subject. We have, then, a rule which embodies a moderately complex arithmetic procedure, and which is *stored* in explicit representational form before being applied. Now, given that humans are clearly *capable* of internally representing explicit, complex arithmetic rules *before* applying them, the question naturally arises whether children commonly learn algorithms such as long multiplication and division by storing explicit representations of these algorithms.

While I have no conclusive answer to this question, the following considerations suggest that children *do* learn complex algorithms by storing explicit representations. First, we should note that, when taught skills such as long multiplication, children are in fact *told* explicit rules (e.g., "after multiplying the entire number by the rightmost digit, move one digit to the left and multiply through again. But this time write down the answer on a new line. But start writing one place to the left . . . "). Of course, these explicit rules are not instantly remembered, and examples are required, but students of average intelligence learn long multiplication after hearing the rules several times and practicing on perhaps 20 or 30 problems. My numbers here are estimates, but surely the number of iterations on which children receive *explicit feedback from a teacher* is seldom above 50. Now, though I know of no existing connectionist simulations of the learning of long multiplication, recent work by Cottrell & Tsung (1989) on the *addition* of 3-digit numbers required on the order of 3000 distinct training examples, and several thousand iterations, to achieve a reasonable degree of generalization (even though back propagation of error was employed, and the numerals were restricted to base four). Given that long multiplication is a considerably more complex task than 3-column addition, and considering that the number of iterations typically required to train c-nets to recognize even simple patterns is well into the hundreds, it would indeed be surprising if c-nets could be trained to perform long multiplication with a sample set of less than several hundred training examples (being very optimistic). In any case, given existing techniques for training c-nets, it is very difficult to believe c-nets could be trained to perform long multiplication when the training examples are restricted to less than 50 patterns.

Now, it might be objected that I have ignored the obvious fact that children receive prior, intensive training in the particular sub-skills involved in long multiplication, and that when the requisite sub-skills (which are already in place) are brought into the learning task, the complexity of the *training* process will be greatly diminished. However, I readily concede that the existence of the requisite sub-skills reduces the complexity involved. That is precisely why I allow that the total number of training patterns may fall into the hundreds. (To adopt an even lower estimate would fly in the face of numerous experiments involving far simpler patterns.) But, this is still an order of magnitude greater than the human learning situation.

Still, a different objection may arise, which runs as follows: "Granted, a human child may be successfully trained in long multiplication using less than 50 patterns, but for all we know each of these patterns is re-presented to the child thousands of times, by some (presently unknown) neural mechanism which works below the level of consciousness. If this possibility is conceded, should we not compare the child's learning task to a system of c-nets which is permitted to process up to 50 distinct patterns, where each pattern is re-presented several thousand times? Once we imagine *this* to be the comparable c-net learning task, is it still obvious that the child will outperform the system of c-nets?"

To reply: very well, if the present suggestion is adopted, then it does become an open question whether a system of c-nets could match the child's performance -- especially if this suggestion is combined with the previous suggestion that learning processes ought to exploit the existence of previously acquired sub-skills. However, there is no physiological evidence to support the conjecture that humans possess neural mechanisms which "re-present the data" to our c-nets many thousands of times. Therefore, rather than accept this purely *ad hoc* suggestion, why not accept what appears to be the obvious solution? That is, children learn long multiplication by remembering the *explicit* rules they are taught, and *by using* these rules to apply the sub-skills they have already acquired to the proper objects in the proper sequence. Examples (1) and (3) have already established that humans *possess* the ability to apply explicit rules in this fashion. Why not suppose that children *use* these abilities in situations where they are provided with explicit rules?

In any case, it seems clear that we ought at least to seriously entertain the possibility that children do exactly this. However, I have discovered no connectionist research which attempts to model the learning of explicit arithmetic algorithms in this fashion. I suggest that the reasons for this are twofold: (a) connectionists are reluctant to *integrate* the classical paradigm (of having explicit representations *control* the sequencing of lower-level functions) into the existing connectionist paradigm, which treats all rules as implicit. (b) This reluctance arises (in part) because of the complexity of the task. It is very difficult to imagine how c-nets can support higher-level, representational control processes without resorting to more conventional (though possibly parallel) architectures. I submit, however, that examples (1) and (3) establish the following: if connectionism is to provide a model for *all* cognitive phenomena, it must include mechanisms for *explicit* rule representation and application. These mechanisms must be general enough to accommodate rules of novel structure. If connectionists can accept and meet the challenge of devising these mechanisms, they will have gone a long way towards integrating the prevailing classical and connectionist paradigms.

6. Summary

We have examined three examples of rule following in which the immediate *representation, and application* of rules appears to *require* the presence of general rule application mechanisms. Two of these examples involve rules of novel structure, which argues for a high degree of flexibility in these application mechanisms. Although the remaining example belongs to the

more *syntactically regular* class of condition-action rules, it should be remembered that one prominent cognitive theory attempts to model most higher-level cognition in terms of (condition-action based) production systems (Anderson, 1976). Moreover, as I have argued, the *existence* of the *kinds* of general rule-application mechanisms considered here strongly suggests that explicit rule representation and rule following are not isolated exceptions, but are important features of human cognition. Furthermore, I have argued that at least *some* explicit rule following is best modelled by the paradigm of classical symbol manipulation. Admittedly, it is uncertain whether *most* high-level processes, such as planning and abstract reasoning, involve explicit rule following and/or symbolic manipulations, but, in light of the fact that *neural mechanisms do in fact sometimes* support these classical processes, we must regard it as a serious open question whether most higher-level cognition involves these *classical* processes. Of course, nothing I have said here would suggest that *all* rule following ought to be modelled on the classical paradigm. Indeed, I have suggested elsewhere (Hadley, 1989) that semantic *grounding* rules are best modelled by connectionist methods. However, I believe the arguments presented here seriously challenge the prevailing connectionist methodology of modelling *all* rules by means of *implicit*, neurally-wired networks. Moreover, our conclusions present the connectionist with a formidable scientific challenge, which is, to show how general purpose rule following mechanisms may be implemented in a connectionist architecture.

References

Anderson, J.R. (1976) *Language, Memory and Thought*, Lawrence Erlbaum Associates, Hillsdale, N.J.

Ajjanagadde, V. & Shastri, L. (1989) "Efficient Inference with Multi-Place Predicates and Variables in a Connectionist System", Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, Ann Arbor, pp. 396-403.

Barnden, J.A. (1988) "The Right of Free Association: Relative-Position Encoding for Connectionist Data Structures", Proceedings of the Tenth Annual Conference of the Cognitive Science Society, Montreal, pp. 503-509.

Cottrell, G.W. & Tsung, F. (1989) "Learning Simple Arithmetic Procedures", Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, Ann Arbor, pp. 58-65.

Elman, J.L. (1989) "Structured Representations and Connectionist Models", Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, Ann Arbor, pp. 17-23.

Fodor, J.A. & Pylyshyn, Z.W. (1988) "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition*, Vol. 28, pp. 3-71.

Hadley, R.F. (1989) "A Default-Oriented Theory of Procedural Semantics", *Cognitive Science*, Vol. 13, pp. 107-138.

Pylyshyn, Z.W. (1984) *Computation and Cognition*, Bradford Books, MIT Press, Cambridge, Ma.

Smolensky, P. (1987) "The Constituent Structure of Mental States: A Reply to Fodor and Pylyshyn", *Southern Journal of Philosophy*, Vol. 26, Supplement, pp. 137-160.

St. John, M.F. & McClelland, J.L. (1989, in press) "Learning and Applying Contextual Constraints in Sentence Comprehension", *Artificial Intelligence*.

Touretzky, D.S. & Hinton, G.E. (1988), "A Distributed Connectionist Production System", *Cognitive Science*, Vol. 12, pp. 423-466.

van Gelder, T. (1989) "Compositionality and the Explanation of Cognitive Processes", Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, Ann Arbor, pp. 34-41.

Centre for Systems Science

• Simon Fraser University established the **Centre for Systems Science** as a provincial centre for advanced study. The objective of CSS is twofold: coordinate and combine advanced systems research at SFU, and encourage application of research results within industry in order to further the technological development of British Columbia.

• CSS has assembled three research groups to streamline its efforts. The orientation of these groups is as follows: intelligent systems, microelectronics, and computer and communications systems. Because of shared research interests, this latter group merges with LCCR (**Laboratory for Computer and Communications Research**) for the publication of technical reports.

• CSS's three groups involve over 80 skilled faculty researchers principally from SFU's Schools of Engineering Science and Computing Science, but also draw substantially on Mathematics, Physics, Geography, Philosophy, Linguistics, Psychology, Communication, and Kinesiology.